心理科学进展 2021, Vol. 29, No. 3, 381–393
Advances in Psychological Science

https://dx.doi.org/10.3724/SP.J.1042.2021.00381

·研究方法(Research Method)·

解读不显著结果：基于500个实证研究的量化分析

王 珺[1] 宋琼雅[1] 许岳培[2,3] 贾彬彬[4] 陆春雷[5] 陈 曦[6]
戴紫旭[7] 黄之玥[8] 李振江[9] 林景希[10] 罗婉莹[11] 施赛男[12]
张莹莹[13] 臧玉峰[14] 左西年[15] 胡传鹏[16]

([1]中山大学心理学系，广州 510006) ([2]中国科学院行为科学重点实验室 (中国科学院心理研究所)，北京 100101) ([3]中国科学院大学心理学系，北京 100049) ([4]上海体育学院心理学院，上海 200438) ([5]浙江师范大学教师教育学院，金华 321000) ([6]个人，上海 200122) ([7]华南师范大学心理学院，广州 510631) ([8]Tisch School of the Arts, New York University, New York 11201, the United States) ([9]苏州大学教育学院，苏州 215123) ([10]黑龙江大学教育科学研究院，哈尔滨 150080) ([11]北京大学心理与认知科学学院，北京 100871) ([12]华东师范大学心理与认知科学学院，上海 200063) ([13]西南大学心理学部，重庆 400715) ([14]杭州师范大学认知与脑疾病研究中心，杭州 311121) ([15]北京师范大学认知神经科学与学习国家重点实验室，北京 100875) ([16]Leibniz Institute for Resilience Research, 55131 Mainz, Germany)

摘 要 不显著结果(如，$p > 0.05$)在心理学研究中十分常见，且容易被误解为接受零假设的证据，并可能导致分组匹配研究的错误推断或者忽视被小样本的不显著结果掩盖的真实效应。但国内目前尚无实证研究对不显著结果的普遍性及其解读进行调查。本研究调查500篇中文心理学实证研究，统计其摘要中出现与不显著结果相关的阴性陈述的频率，判断并统计基于阴性陈述的推断准确性，并使用贝叶斯因子对不显著结果中包含t值的研究进行重新评估。结果表明，36%的摘要提及不显著结果，共包含236个阴性陈述。其中，41%的阴性陈述对不显著结果的解读出现偏差(如，解读为支持了零假设)。对包含t值的研究进行贝叶斯因子分析，结果显示仅有5.1%的不显著结果可以提供强证据支持零假设($BF_{01} > 10$)。与先前对国际心理学期刊的调查结果相比(32%的摘要包含阴性陈述；72%的阴性陈述对不显著结果的解读错误)，中文心理学期刊中报告不显著结果的比例更高，且对不显著结果解读错误的比例更低。但国内研究者仍需进一步加强对不显著结果的认识，推广适于评估不显著结果的统计方法。

关键词 不显著结果，零假设显著性检验，贝叶斯因子，元研究
分类号 B841

1 引言

基于正确的统计推断方法是从数据中得到正确结论的重要前提之一。当前科研实践中，主导的统计推断方法是零假设显著性检验(Null hypothesis significance testing, NHST) (American Psychological Association, 2010; Wasserstein & Lazar, 2016)。在此框架下，研究者通常根据p值大小做出是否拒绝零假设的二分决策。具体而言，当p值小于某个预设的α阈值时(通常设为0.05)，研究者可以拒绝零假设从而接受备择假设；而当p值大于此阈值时，研究者无法拒绝零假设。然而，无法拒绝零假设存在两种可能：一是数据支持零假设，即效应不存在(evidence of absence)；二是缺乏充分的统计功效，因而未检测到真实存在的效应(Dienes, 2014, 2016)，即没有证据表明效应存在(absence of evidence)。

研究者很早就意识到NHST的局限性(Amrhein et al., 2019; Edwards et al., 1963; Gigerenzer et al.,

收稿日期：2020-07-14
通信作者：胡传鹏, E-mail: hcp4715@hotmail.com

2004; Meehl, 1967; Miller, 2011; Nickerson, 2000; Ziliak & McCloskey, 2008)。一方面，基于 NHST 的二分决策思维一定程度上导致了研究者对不显著结果的忽视甚至歧视，进而引发了论文发表过程中的发表偏倚(publication bias)。Fanelli (2012)分析各学科的文献后发现，各种学科发表的论文中，阳性/显著结果的比例均大于阴性/不显著结果的比例，而心理学论文中阳性结果的比例高达95%以上。这种发表偏倚可能会导致研究者对真实效应的错误估计(Algermissen & Mehler, 2018; Schäfer & Schwarz, 2019)，从而在一定程度上加剧了心理学领域的可重复性危机(Baker, 2016; Ioannidis, 2005; Klein et al., 2014; Open Science Collaboration, 2015; 胡传鹏 等, 2016)。另一方面，研究中还会出现对不显著结果的错误解读，即尽管 $p > 0.05$ 无法区分“数据支持零假设”与“数据不足以支持或拒绝原假设”这两种情况，但研究者在结论表述中经常出现混淆，错误地将 $p > 0.05$ 作为支持零假设的证据，影响结论的可信度(Greenland et al., 2016; X. Lyu et al., 2020; Z. Lyu et al., 2018; 胡传鹏 等, 2016; 骆大森, 2017)。Lyu 等人(2020)调查发现，53%的研究者错误地认为，当 $p > 0.05$ 时，数据支持了零假设。

上述对不显著结果的错误解读可能带来两个严重的后果。第一，错误地接受零假设会影响对随后干预效果的推断。在临床试验中，研究者多使用卡方检验或独立样本 t 检验来分析实验组与控制组在一些混淆变量上的差异(例如性别、年龄、教育程度)。当 t 检验的 p 值大于 0.05 时(如 0.06)，研究者可能认为两组在该变量上没有差异，在后续对干预效果的分析中不再考虑该变量的影响，忽视了该变量可能带来的严重混淆。第二个后果是对阴性结果的忽视。研究者可能由于小样本等原因缺乏足够的检验力来检测到实际存在的效应，进而得到不显著结果(Button et al., 2013; Chen et al., 2018)。在这种情况下，如果错误解读不显著结果，研究者会得出效应不存在的结论，这样可能错失潜在重要的效应(Fiedler et al., 2012)。例如，一项多中心合作的元分析(meta-analysis, 也译为荟萃分析)显示，尽管帕金森病患者的左侧壳核在元分析结果中是全脑最异常的脑区；但单个中心的结果中，由于检验力较低，只有 2 个中心的壳核异常在进行多重比较矫正后仍达到显著水平(Jia et al., 2018)。

尽管目前对于 NHST 框架下不显著结果的讨论逐渐增多(吕小康, 2014; 仲晓波, 2016)，但是大多数是基于理论与方法的探讨，缺乏实证性研究探讨当前国内心理学领域发表论文中不显著结果的普遍性和解读情况。Aczel 等人(2018)回顾 2015 年发表在 *Psychonomic Bulletin & Review, Journal of Experimental Psychology: General* 和 *Psychological Science* 上的 412 篇实证研究论文，发现摘要中包含阴性陈述(研究者直接阐明效应不存在或者提及不显著的结果)的文章接近 1/3，这其中有 72%的文章都存在对不显著结果的错误解读。那么，在国内心理学领域的权威期刊中，是否也存在类似的错误解读不显著结果的情况？

此外，在实际研究中，研究者有时确实需要证实零效应或者零假设为真。如前所述，对于被试间设计的组间匹配问题，研究者需要尽量保持实验组与控制组在某些属性上(如年龄、性别)的一致性。在这种情况下，研究者需要能够为“其他方面没有差异”这个零假设提供证据。有时，研究者也有可能需要检验两个相互竞争的理论，用实验数据说明其中一个理论所预测的差异并不存在，即支持零假设。换言之，在某些情境下，证实“零假设为真”才是研究者所想要达到的目标。这一目标服务于拒绝或者证否某个研究假设、提出替代的新研究假设，从而促进科学理论的发展。

由于 NHST 无法为零假设提供支持，而用 $p > 0.05$ 为零假设提供支持实际上是错误的做法(Chuard et al., 2019)。因此，研究者需要引入合适的统计方法探究数据支持零假设的程度，如贝叶斯因子(Bayes factors, BFs) (Wagenmakers et al., 2018; Wagenmakers et al., 2011; 胡传鹏 等, 2018)。Aczel 等人(2018)对基于 t 检验得出不显著结果的数据进一步计算其贝叶斯因子，进而评估数据支持零假设的程度，结果表明在这些不显著的 t 检验结果中，只有 3%的 t 检验能够得到较强的证据支持零假设($BF_{01} > 10$)，71%的 t 检验仅能够得到中等强度的证据支持零假设($10 > BF_{01} > 3$)。这一结果表明，在缺乏恰当的统计方法的情况下，研究者可能忽视了一个重要的问题，即得到不显著结果的数据无法提供足够强的证据支持零假设。这一研究现象在国内心理学核心期刊已发表论文中是否存在，也是值得探索的问题。理解这一

现象能进一步帮助国内心理学研究者了解到, 错误解读不显著结果会得到错误支持零假设的结论。

为了了解中文心理学研究论文中对于不显著结果的解读现状, 本研究参考 Aczel 等人(2018)的文章, 调查了 5 本国内心理学的核心期刊(《心理学报》、《心理科学》、《中国临床心理学杂志》、《心理发展与教育》以及《心理与行为研究》)在 2017 年与 2018 年发表的实证研究论文。具体而言, 本研究分析了随机抽取的 500 篇论文中不显著结果的报告情况和错误解读的比例, 通过计算贝叶斯因子评估得到不显著结果的数据是否确实可以支持零假设, 并评估其支持的程度。之后, 本研究还对比了中文核心期刊和国际期刊中对于不显著结果解读现状的差异。本文旨在帮助研究者意识到对不显著结果出现误读的普遍性, 进而在统计推断过程中更为谨慎细致, 避免错误解读的发生。

2　方法

2.1　文章抽样

本研究选取 5 本可以免费下载全文的国内心理学核心期刊, 分别是《心理学报》、《心理科学》、《中国临床心理学杂志》、《心理发展与教育》以及《心理与行为研究》。这些期刊涵盖了不同领域的心理学研究, 如认知心理学、发展心理学、社会心理学、临床心理学等。随后, 整理出这 5 个杂志于 2017~2018 年发表的所有实证研究论文, 即包含数据分析部分的论文(不包括综述、元分析或者评论等), 摘录各个杂志中 2017~2018 年所有实证研究论文的标题、出版时间、卷号、页码, 并为每篇文章进行编号。例如, 《心理学报》的第 2 篇文献编码为 1002——1 表示心理学报所对应的杂志 ID (不同杂志对应不同的杂志 ID), 002 表示该文献是在杂志中的排序。具体编码规则见 https://osf.io/mf42q/。最后, 根据每个期刊发文量, 按比例对每个期刊的实证研究论文进行随机抽取。《心理学报》、《心理科学》、《中国临床心理学杂志》、《心理发展与教育》以及《心理与行为研究》的实证研究数目分别为 246、299、379、162、213, 总共 1299 篇文章, 对应的发文比例分别为 18.94%、23.02%、29.18%、12.47%、16.40%。因此, 随机抽取的文章数目分别为: 《心理学报》95 篇、《心理科学》115 篇、《中国临床心理学杂志》146 篇、《心理发展与教育》62 篇、《心理与行为研究》82 篇。用于随机抽取文章的代码见 https://osf.io/7my4g/。

2.2　文章编码

编码过程分为 3 步, 分别是初步编码、编码校对和分类编码及校对(图 1)。在初步编码中, 我们将选择的 500 篇文献随机分为 13 份, 分配给 13 名编码人员。具体编码过程如下: 阅读每篇文章的摘要, 判断其是否包含至少一个阴性陈述(Negative statement, 也被译为负性陈述, 两者为同一概念, 本文统一使用阴性陈述)。“阴性陈述”是

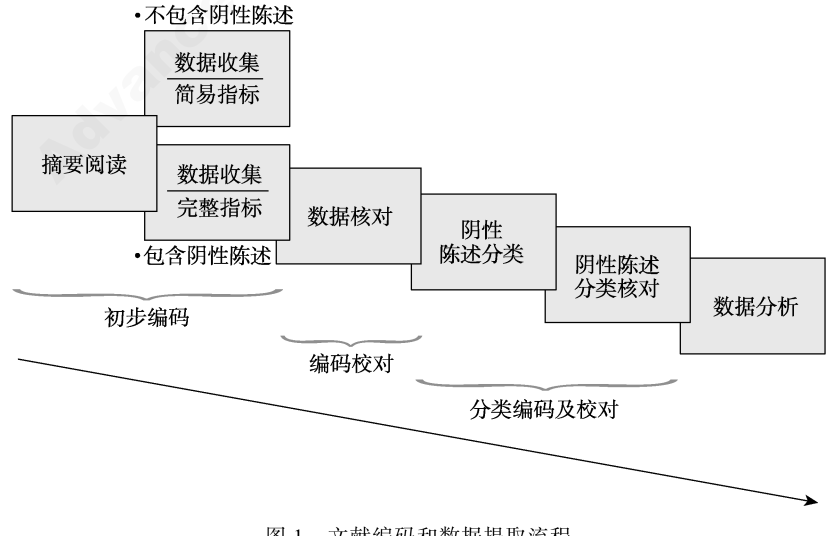


图 1　文献编码和数据提取流程

指研究者直接阐明效应不存在(如“干预组和控制组之间不存在差异”), 或者提及不显著的结果(如“没有证据支持干预组和控制组有显著差异”)。如果摘要中不包含阴性陈述, 那么编码人员只需要摘录文章的基本信息, 包括文章编号、引用、文章链接以及文章类型。如果摘要中包含至少一个阴性陈述, 那么除了以上基本信息之外, 还应摘录该阴性陈述以及正文中与其对应的统计检验信息。其中, 统计检验信息主要包括统计检验方法, 当统计检验方法为 *t* 检验时(包括单样本 *t* 检验, 配对样本 *t* 检验和独立样本 *t* 检验), 还需要摘录 *t* 值, *p* 值和样本量。这部分信息用于后续的贝叶斯因子计算。

为确保编码内容的准确性, 在完成初次编码之后, 重新分配文章, 进行编码校对工作。具体的编码模板和编码流程可以参考补充材料(https://osf.io/a39hb/)。

得到文章的阴性陈述及相应的统计结果数据后, 先由 6 名编码人员独立进行阴性陈述的分类编码, 随后共同讨论存在分歧的分类编码结果, 得到最终的阴性陈述分类结果。具体类别及其分类标准见表 1。为了评估 6 名评分者的一致性, 使用 Gamer 等人(2019)开发的 R 包 *irr* (函数 *kappam.fleiss*)计算了 Fleiss' kappa (Fleiss, 1971)。该指标适用于编码变量为分类变量且评分者多于两位的情况。

2.3 贝叶斯因子分析

为了重新评估采用 *t* 检验(单样本 *t* 检验、配对样本 *t* 检验或独立样本 *t* 检验)的研究数据支持零假设的程度, 我们根据文章报告的统计检验参数(样本量和 *t* 值)计算贝叶斯因子(Ly et al., 2018)。贝叶斯因子可以用于比较数据支持备择假设(H_1)和零假设(H_0)的相对程度(Wagenmakers et al., 2018), 公式如下:

$$\mathrm{BF}_{01}=\frac{P\left(Data|H_0\right)}{P\left(Data|H_1\right)}$$

BF_{01} 的下标 1 表示 H_1, 0 表示 H_0。因此, BF_{01} 代表 H_0 与 H_1 对比的贝叶斯因子, 而 BF_{10} 代表 H_1 与 H_0 对比的贝叶斯因子。例如, $\mathrm{BF}_{01} = 10$ 表示在零假设 H_0 为真的条件下出现当前数据的概率是备择假设 H_1 为真的情况下出现当前数据概率的 10 倍。基于 Jeffreys (1961)对于不同 BF_{01} 值对应意义的划分, Wagenmakers 等人(2018)明确了不同大小的 BF_{01} 对应的意义。然而, 这种划分方式仅作参考, 研究者需要根据特定的研究问题对 BF_{01} 的意义进行评估。

参考 Aczel 等人(2018)的研究, 使用 Morey 等(2015)开发的 R 包 *BayesFactos* (函数 *ttest.tstat*)计算 BF_{01}。该软件包的默认设置是使用双尾柯西分布(Cauchy distribution)作为备择假设的先验($r=\frac{\sqrt{2}}{2}$, r 为尺度参数, 也有文献中使用 γ)。先前研究表明这种备择假设的先验设置是比较恰当的(Ly et al., 2016a, 2016b; Rouder et al., 2009)。同时, 为了探究贝叶斯因子结果的稳定性, 我们选择不同的先验分布分别计算贝叶斯因子。其中一种先验分布为正态分布(Dienes, 2014), 相比于默认先验, 正态先验分布在 0 附近的概率密度相对更大, 因此得到的效应比默认先验的结果更接近 0。另一种先验分布为 Gronau 等人(2019)基于专家意见确定的效应量分布(即有信息的先验), 反映了专家对于效应量分布的信念(中位数为 0.350)。

表 1 阴性陈述的具体类别以及分类标准

类别	分类标准	示例
基于频率主义的正确解读	根据 NHST 的逻辑对不显著结果进行解读, 即仅说明其结果无法拒绝零假设, 或无法支持备择假设。	结果表明没有证据支持干预组和控制组有(显著)差异。
基于频率主义的错误解读——推广至总体	将不显著结果解读为支持了研究中样本所在总体水平上的零假设。	结果表明干预没有效果。
基于频率主义的错误解读——基于当前样本	将不显著结果解读为支持了研究中样本中的零假设。	结果表明干预组和控制组之间没有差异。
基于贝叶斯因子的解读	利用贝叶斯因子支持零假设而非备择假设。	$\mathrm{BF}_{01} > 10$, 表明有强的证据支持零假设。
难以判断	由于阴性陈述的语言措辞, 对其类别难以做出明确判断。	除恐惧情绪外, 基本表情的强度越大, 被试对表情的识别越好。

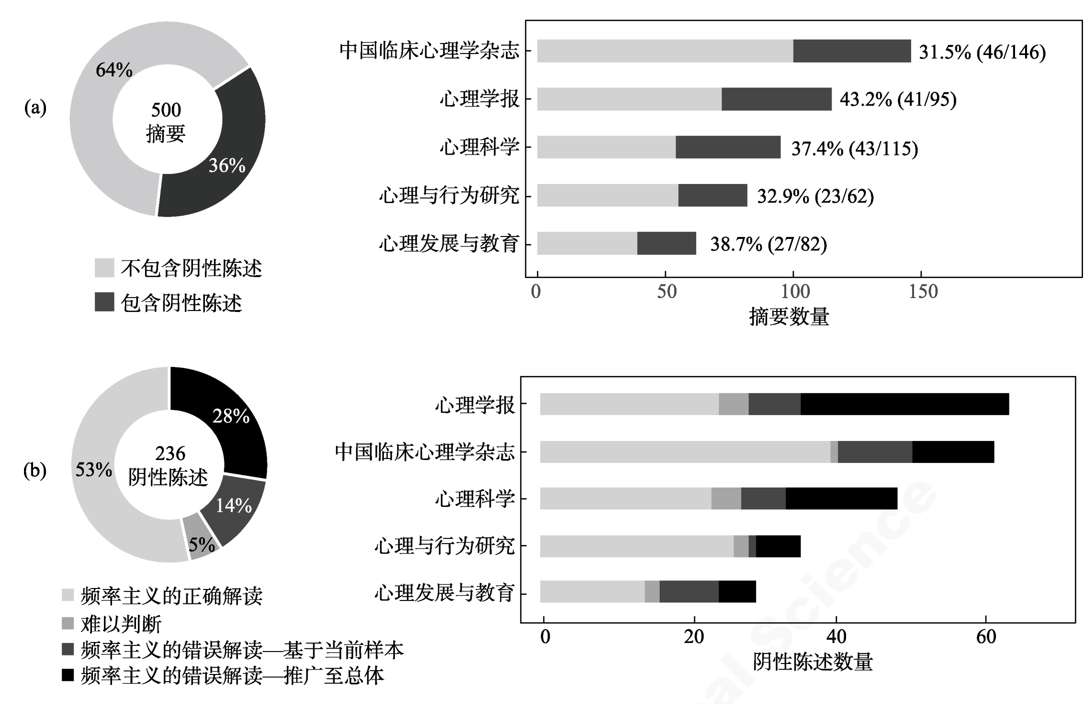


图 2　(a)阴性陈述在不同杂志中的占比; (b)阴性陈述的解读分类在不同杂志中的占比
注：此分类是基于解读①, 见正文关于两种解读的说明。

考虑到研究者可能误把 p 值作为支持零假设的证据, 我们进一步探究了 p 值与 BF_{01} 之间的关系, 即计算了 p 值与 BF_{01} 之间的相关系数 τ (Kendall's τs) (Kendall & Gibbons, 1990)及其对应的 95%可信区间(credible intervals, CIs), 以此评估 p 值是否与 BF_{01} 存在较强的相关。如果 p 值与 BF_{01} 存在较强的相关, 则较大的 p 值从某种程度上可以支持零假设; 假如 p 值与 BF_{01} 不存在较强的相关, 尤其是当 $p > 0.05$ 时与 BF_{01} 的相关较弱, 则表明使用较大 p 值作为支持零假设的证据是错误。由于所分析的相关关系并非线性关系, 我们选择使用相关系数 τ。我们使用 Signorell (2017)开发的 R 包 DescTools 中的函数 *KendallTauB* 计算 τ; 基于 τ 和 t 检验数目通过函数 *credibleInterval KendallTau* (van Doorn et al., 2018)计算对应的 95% CIs。最后, 由于大样本量的研究往往能够提供更强的证据, 我们也采用同样的方法探索了 BF_{01} 与样本量之间的相关关系。

3　结果

3.1　不显著结果在中文文献中的普遍性

本次分析的结果发现, 在 500 篇实证文章中, 36%的文章摘要包含了至少一个阴性陈述。其中《心理学报》上发表的文章摘要包含阴性陈述的比例最高(43%), 但是所有杂志的这一比例都超过了 30% (见图 2a)[1]。这一结果表明, 在心理学实证研究论文中阴性陈述非常普遍。

[1] 由于 500 篇文章的研究类型包括实验、准实验和问卷调查, 那么不同的研究类型中出现阴性陈述的比例可能不同。因此我们分析了在不同杂志中, 不同研究类型下阴性陈述分布情况。结果发现, 实验(45.8%)和准实验(36.2%)在阴性陈述中占比相对问卷调查(17.9%)较大。但是由于不同的杂志研究侧重的研究方向不同, 使得不同研究类型在不同杂志中的占比差异较大。例如在《心理学报》、《心理科学》以及《心理与行为研究》中, 实验类的文章占比较大, 均超过 50%。而在《中国临床心理学杂志》和《心理发展与教育》中, 问卷调查和准实验研究的比例较大。需要强调的是, 同一篇文章可能包含多个研究, 因此我们同时考虑了同一个研究对应不同的阴性陈述(例如, 阴性陈述 1: 研究结果没有发现研究 1 中的变量 A 对于反应时有显著影响; 阴性陈述 2: 研究结果没有发现研究 1 中的变量 B 对于反应时有显著影响)以及同一个阴性陈述对应不同研究的情况(例如, 在阴性陈述研究 1 和研究 2 中都没有发现变量 A 对于反应时有显著影响)。因此考虑研究类型的阴性陈述的总数为 301, 超过前文提到的 236 个阴性陈述。

3.2 阴性陈述分类

对于 6 位评分者关于阴性陈述分类的一致性分析结果表明, Fleiss′ kappa 为 0.588 ($p < 0.001$)。参考 Landis 和 Koch (1977)对 Fleiss′ kappa 含义的划分, 该 Fleiss’ kappa 表示中等强度的评分者一致性。此外, 作者共同讨论了存在分歧的分类编码结果, 进而得到最终的阴性陈述分类结果。因此, 阴性陈述分类结果较为可靠。

在分类过程中发现, 阴性陈述中常出现类似于“没有显著的差异/效应/作用”的描述(n = 55)。由于汉语表达的模糊性, 对此类陈述可以有两种解读: 解读①认为此类陈述是对 p < 0.05 的直接解读, 即“差异没有达到统计上显著水平”, 分类为“基于频率主义的正确解读”; 解读②则认为此类陈述是支持零假设的描述, 等同为“没有差异/效应/作用”, 即分类为“基于频率主义的错误解读−基于当前样本”。因此, 在后续对阴性陈述的分类结果的描述中, 我们分别依据这两种解读进行了说明。

我们将“没有显著的差异/效应/作用”的陈述分类为基于频率主义的正确解读, 对 236 个阴性陈述进行分类。结果显示, 基于频率主义的正确解读占 53.4% (n = 126); 基于频率主义的错误解读占 41.1% (n = 97), 其中 13.6% (n = 32)落在子类别基于频率主义的错误解读−基于当前样本中, 27.5% (n = 65)落在子类别基于频率主义的错误解读−推广至总体中。此外还有 5.5% (n = 13)阴性陈述表述不清晰, 难以明确具体的阴性陈述类别, 故编码为“难以判断”。具体各类别在各个杂志中的分布见图 2b。基于解读②的分类结果见脚注 [2]。Aczel 等人 2018 年的研究中还考虑了基于贝叶斯分析的阴性陈述类别, 但是我们并未发现从属于此类别下的阴性陈述, 即在这些文献中并没有使用贝叶斯因子评估支持零假设的程度的案例, 因此在本研究中剔除该类别。

3.3 贝叶斯因子分析

在 NHST 框架下, 研究者只能根据 p 值大小做出是否拒绝零假设的二元决策, 因而无法得到支持零假设的证据。因此, 我们结合 t 检验的数据重新计算 BF_{01}, 进而评估得到不显著结果的数据支持零假设的程度。

在所有统计检验中, 使用 t 检验且报告了 t 检验统计量和样本量的统计检验数目为 39。根据 t 检验的 t 值和样本量, 使用中等尺度的双尾柯西分布(Cauchy distribution)作为备择假设的先验计算 BF_{01}, 范围在 0.51 到 10.64。参考 Wagenmakers 等人(2018)对 BF_{01} 含义的划分, 以 1、3 和 10 为临界值将 BF_{01} 划分为“较弱的证据支持 H_1”, “较弱的证据支持 H_0”, “中等程度的证据支持 H_0”和“强的证据支持 H_0”。结果表明, 39 个 t 检验中有 2.6% (n = 1)的 BF_{01} 表明有较弱的证据支持 H_1, 33.3% (n = 13)的 BF_{01} 表明有较弱的证据支持 H_0, 59% (n = 23)的 BF_{01} 表明有中等程度的证据支持 H_0, 而只有 5.1% (n = 2)的 BF_{01} 表明有强的证据支持 H_0。换言之, 如果作者在原文中做出了支持 H_0 的推断, 则 BF_{01} 表明这些检验中只有一半左右有中等或强的证据支持 H_0。因此, 研究者基于 p 值推断 H_0 为真是不恰当的。

为了验证结果的稳健性, 避免先验设定对结果造成影响, 我们分别使用正态先验和有信息先验重新计算贝叶斯因子。不同先验设置下 BF_{01} 的分布如图 3a 所示。基于正态先验, BF_{01} 的范围为 0.45 到 6.00; 其中有 15.4% (n = 6)的 BF_{01} 表明有较弱的证据支持 H_1, 64.1% (n = 25)的 BF_{01} 表明有较弱的证据支持 H_0, 20.5% (n = 8)的 BF_{01} 表明有中等程度的证据支持 H_0。而基于有信息先验, BF_{01} 范围为 0.41 到 21.69; 其中 20.5% (n = 8)的 BF_{01} 表明有较弱的证据支持 H_1, 53.8% (n = 21)的 BF_{01} 表明有较弱的证据支持 H_0, 17.9% (n = 7)的 BF_{01} 表明有中等程度的证据支持 H_0, 而只有 7.7% (n = 3)的 BF_{01} 表明有强的证据支持 H_0。由此可见, 基于不同的先验设定, BF_{01} 的分布存在差异。

研究进一步探究先验设置对于阴性陈述分类的影响。结果表明, 将默认先验更改为有信息先验时, BF_{01} 所对应的含义发生更改的比例为 60% (n = 23); 而将默认先验更改为正态先验时, BF_{01}

---

[2] 如果我们将“没有显著的差异/效应/作用”的陈述分类为基于频率主义的错误解读−基于当前样本, 再次对 236 个阴性陈述进行分类。解读的改变只影响基于频率主义的正确解读和基于频率主义的错误解−基于当前样本这两个类别的阴性陈述数目, 不影响其余两类的陈述分类。结果显示, 基于频率主义的正确解读占 30.1% (n = 71); 相对的, 基于频率主义的错误解读占 64.4% (n = 152), 其中 36.9% (n = 87)落在子类别基于频率主义的错误解读−基于当前样本中, 27.5% (n = 65)落在子类别基于频率主义的错误解读——推广至总体中。

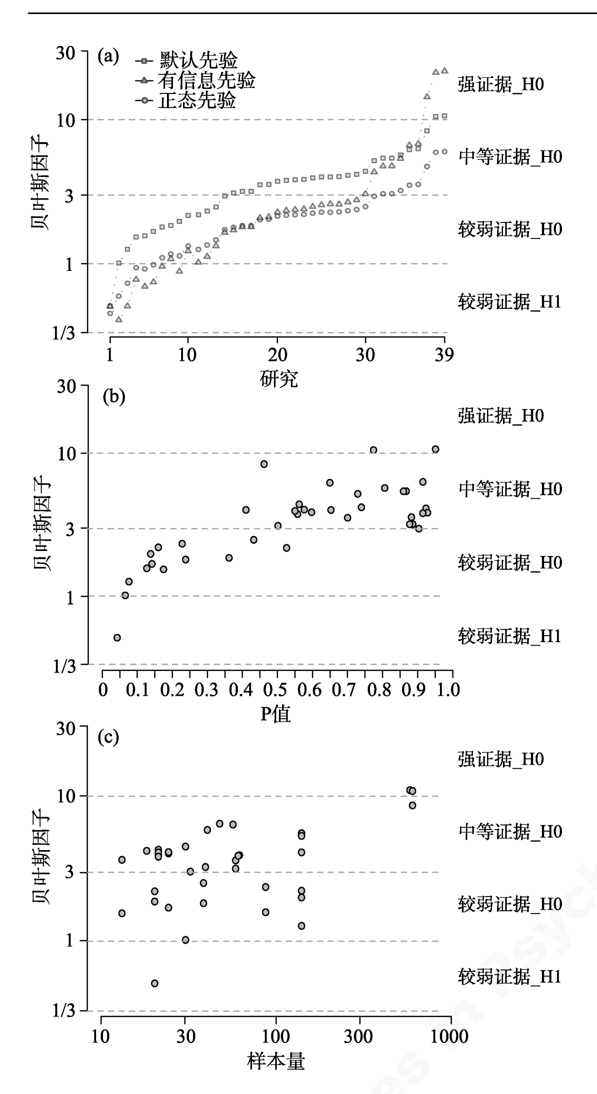


图 3　(a)不同先验设置下 BF_{01} 的分布及含义; (b)默认先验下的 BF_{01} 与 p 值的关系; (c)默认先验下的 BF_{01} 与样本量的关系

注: 针对同一个样本可能存在多个 BF_{01} 值, 例如样本量为 138 的样本对应多个 BF_{01}。

所对应的含义发生变化的比例为 61.5% ($n = 24$)。这说明先验分布的设置对于 BF_{01} 所对应的含义有较大影响, 研究者在计算 BF_{01} 时需要谨慎选择恰当的先验分布。

最后, 我们对于贝叶斯因子进行了探索性的分析, 分别探究 BF_{01} 与 p 值以及 BF_{01} 与样本量的相关关系。由于使用 t 检验且报告了 t 检验统计量和样本量的统计检验数目仅为 39, 本研究的相关分析结果仅供参考, 需要后续研究进一步验证结果的可靠性。为了探究 p 值与对应的 BF_{01} 的相关关系, 我们绘制了 p 值与 BF_{01} 的散点图(图 3b), 并计算了相关系数 τ 及其对应 95%可信区间。结果显示, p 值与 BF_{01} 的相关系数 τ 为 0.527, 其 95% CI 为[0.282, 0.687]; 说明 p 值越大, 对应的 BF_{01} 值也越大。然而, 从图 3b 可以看出, 相关关系主要体现在 p 值较小($p < 0.2$)的不显著结果当中; 随着 p 值增大, BF_{01} 的变化趋近平缓。因此, 该结论的合理性需要后续研究进行探讨。

同样, 研究分析了样本量与 BF_{01} 之间的关系(图 3c)。结果表明, 样本量与 BF_{01} 之间的相关系数 τ 为 0.243, 95%可信区间为[0.018, 0.431], 说明 BF_{01} 与样本量的相关不强。由图 3c 同样可以看出, 随着样本量的增加, BF_{01} 的变化并不明显。然而, 样本量的范围主要在 13 到 138 之间, 只有个别研究样本量超过 300。因此, 结论的准确性还有待进一步验证。

4　讨论

本研究对 500 篇随机选取的中文心理学实证研究论文进行分析, 摘录所有出现在摘要部分的阴性陈述, 并且摘取与阴性陈述相关的统计量、样本量重新计算贝叶斯因子, 旨在探究心理学中文核心期刊中实证论文不显著结果的普遍性及其解读的正确程度, 并与国际期刊的现状进行比较。

在阴性陈述出现的普遍性上, 我们发现有 36%的论文摘要(n = 180)中包含了阴性陈述, 即研究者直接阐明效应不存在或者提及不显著的结果。例如, 对于发表在《心理学报》上的实证研究论文, 摘要中包含阴性陈述的比例超过了 40%。而 Aczel 等人(2018)回顾了发表在国际核心期刊上(*Psychonomic Bulletin & Review, Journal of Experimental Psychology: General* 和 *Psychological Science*)的实证研究后, 发现在摘要部分提及阴性陈述的文章比例为 32%; 这一比例低于我们对国内期刊进行调查统计后得到的结果。结合 Aczel 等人(2018)的研究, 本研究的结果表明不显著结果在心理学研究中不可或缺, 研究者需要不显著结果来辅助其进行推断; 尤其是在实验研究中(即 Aczel 等人(2018)所分析的主要研究类型), 阴性陈述出现的比例高达 45.8%。

在对不显著结果的解读正确率上, 虽然 41.1%陈述中存在对不显著结果的错误解读(将“无显著

差异”分类为基于频率主义的正确解读，即解读①)，但 Aczel 等人(2018)的结果表明国际期刊中错误解读不显著结果的比例高达 72%。即使我们将国内研究者常用的表述“无显著差异”作为错误解读(解读②：64.4%)，错误解读的比例仍然低于国际期刊中的比例。这一结果表明，虽然国内研究者与国际同行相似，对不显著结果的错误解读十分普遍，但表现在文章中的错误解读比例仍然低于国际心理学期刊中的比例。值得注意的是，基于不同解读的分类结果相差 20%以上，这提示研究者需要对此类涉及统计推断的表述进行明确清晰的表达。

此外，贝叶斯因子分析的结果发现，即便考虑了不同先验分布的设置情况，鲜有 BF_{01} 取值大于 10 (默认先验：$n = 2$；正态先验：$n = 0$；有信息先验：$n = 3$)，大部分 BF_{01} 取值小于 3 (默认先验：$n = 14$；正态先验：$n = 31$；有信息先验：$n = 29$)。虽然研究者对于 BF_{01} 所表示的证据强度的解读可能存在差异(Schönbrodt, 2015)，但是大部分还是将 $BF_{01} < 3$ 解读为微弱的支持零假设的证据，将 $BF_{01} > 10$ 解读为强的支持零假设的证据(Lee & Wagenmakers, 2014)。这一贝叶斯因子分析结果与国际期刊的结果较为相似，但由于能够提供较强支持证据的样本量小，所以无法说明在这一点上国内期刊的优势是明显的。贝叶斯因子分析的结果表明，基于得到不显著结果的数据计算获得的 BF_{01} 几乎无法得到强的支持零假设的证据。但是，贝叶斯因子分析中 t 检验对应的样本量大部分小于 100，Aczel 等人(2018)也提出该结果部分原因可能在于心理学研究中的样本量小(Button et al., 2013; Stussi et al., 2018; 崔诣晨 等, 2019; 谢书书 等, 2019)。而 Hoekstra 等人(2018)重新分析了医学领域的不显著结果，发现当样本量大时，数据得到的支持零假设的程度强。

同时，我们通过相关分析探究 p 值大小和样本量大小与 BF_{01} 的相关关系。然而本研究对于 BF_{01} 与 p 值和样本量的相关分析仅仅是简单的探索，同时相关分析涉及的 t 检验数目仅为 39，因此，我们希望有研究可以进一步详细探讨这些变量间的关系，得到更可靠的结论。对于 p 值和 BF_{01}，相关系数为 0.527。但与 Aczel 等(2018)的研究相似的是，我们同样发现 p 值和 BF_{01} 的正相关主要出现在 p 值较小的不显著结果当中。当 $p < 0.2$ 时，BF_{01} 会随着 p 值的增大而增大；但是当 p 值较大时，p 值的增大并不会对 BF_{01} 造成较大的影响。这也反应了 NHST 的局限，即 p 值的大小并没有明确的含义，不能衡量研究假设为真或为假的概率，更大的 p 值并不意味着有更强的证据支持零假设(郝丽 等, 2016; X. Lyu et al., 2020)。Wetzels 等人(2011)的结果也同样表明，当 p 值较大时，BF_{01} 随 p 值的变化幅度小。除了心理学研究，Hoekstra 等人(2018)对医学研究中出现的不显著结果进行分析后，发现 BF_{01} 的 log 形式与 p 值存在线性相关，即随着 p 值的增加，BF_{01} 随 p 值的变化幅度变小。对于样本量与 BF_{01}，相关系数仅为 0.243，说明随着样本量的增加，BF_{01} 的变化幅度小。而 p 值会受到样本量的影响(程开明, 李泗娥, 2019)。即使效应量很小，当样本量足够大时，也很容易得到显著结果。因此，研究结论不应该只关注统计结果是否显著，而是将统计结果与效应的实际意义相结合。不过正如前文所述，贝叶斯因子分析中的 t 检验数目以及对应样本量都较小，因此本结果的普适性有待考证。

值得注意的是，Aczel 等人(2018)发现了有 10%的阴性陈述是基于贝叶斯因子进行统计推断，而非基于 NHST 进行统计推断。而本研究随机选取的 500 篇文章中并没有涉及贝叶斯因子的使用，这在一定程度上反映出国内研究者较少了解能够支持零假设的方法。因此，吕小康(2012)建议研究者需要更多地关注其他统计推断方法作为 NHST 的补充，以适当的难度向研究者介绍不同统计方法背后的原理，从而更全面的了解不同方法的优劣势；例如等价性检验(Equivalence test) (Lakens et al., 2018; Lakens et al., 2018; Rogers et al., 1993)，贝叶斯估计(Bayesian estimation) (Kruschke, 2011; Kruschke & Liddell, 2018; McElreath, 2018)和贝叶斯因子(Bayes factor) (Wagenmakers et al., 2018; Wagenmakers et al., 2011; 胡传鹏 等, 2018)。具体的方法使用可以参考陆春雷等(2020)。

本研究与 Aczel 等(2018)的研究还存在一个重要的区别：本研究的阴性陈述分类额外考虑了“难以判断”的类别。例如，编号为 2052 的文章对于不显著结果的表述为“泛化法任务中，疼痛表情仅在秒上条件延长了主观时距”，这种表述隐含有其他情况下效应不存在或者没有发现其他情况下效应存在的意思，这分别对应着错误解读和

正确解读两种情况。然而我们无法确定作者希望表达的含义, 因此将这种描述分类为“难以判断”。这类含糊的表述在一定程度上反映了研究者对不显著结果表述准确性的忽视, 过分关注显著结果的陈述。此外, 我们还发现文献中用词不规范的情况。例如, 有文章写道“在运动员群体中, 高状态焦虑对加工效能和正确率都影响不大”, 这同样说明研究者需要更加谨慎地对待不显著结果。

本文虽然揭示了当前文献中存在着对不显著结果的错误解读, 但无法对产生这些误解的原因进行探讨。其中一个可能的原因是教科书中关于 p 值的解读存在错误。例如, Cassidy 等人(2019)统计了北美心理学教材关于 p 值的解读, 发现很大一部分教科书对 p 值存在误解。而国内教科书也存在对于不显著结果的错误解读。例如, 张厚粲和徐建平(2015)在第八章写道“假设检验的问题, 就是要判断虚无假设 H_0 是否正确, 决定接受还是拒绝(reject)虚无假设 H_0”; 卢淑华(2009)在第七章写道“如果在原假设 H_0 成立的条件下, 根据样本所计算的某个统计量, 发生的可能性不是很小的话, 那么就接受原假设”。这些表述都认为基于 NHST 可以得到接受零假设的证据。教科书中此类的错误解读可能是国内研究者错误解读不显著结果的原因之一。

本研究也存在几点局限。第一, 负责编码的研究人员共有 13 名, 可能对编码手册的理解存在差异, 例如摘取的阴性陈述篇幅长短不一致。为了减小这些差异的影响, 每篇文章的编码都至少由两位编码者进行编码, 由第二位编码者校对第一位编码者的工作。同时, 对于研究关注的阴性陈述的分类编码, 先由 6 名编码人员独立完成, 再共同讨论存在分歧的编码结果, 并通过 Fleiss' kappa 评估分类结果的评分者一致性, 说明编码结果较为可靠。第二, 本研究通过贝叶斯因子量化数据支持零假设的程度时, 仅使用了 t 检验的数据, 因此许多使用相关分析等其他统计分析方法的数据并未包含在贝叶斯因子计算之中。但是本研究的结果与 Aczel 等人(2017)的结果模式一致。他们对于 35515 篇已发表的文章中出现的基于 t 检验, F 检验和相关分析的显著结果重新计算了贝叶斯因子, 结果发现心理学研究中不同的统计检验得到的证据强度是类似的, 因此本研究中基于 t 检验的数据在一定程度上可以推广到其他的统计检验中。第三, 研究仅统计了 2017 年和 2018 年的数据, 仅能在一定程度上反映当时的情况, 对于近 5 年或者近 10 年情况以及变化趋势可能无法提供数据信息。第四, 在临床试验中, 可能错误地接受零假设进而推断两组在某些变量上是匹配的, 但这些详细的信息一般不出现在摘要中, 将来可以针对该问题进行全文搜索。

虽然本研究存在一些局限, 但是研究结果依然提示心理学乃至其他实证科学的研究者在研究中需要重新审视不显著结果对应的结论。对不显著结果的错误解读可能会带来严重的后果: 忽略了被试间设计中实验组与控制组存在的实际差异; 忽视小样本研究中不显著结果可能掩盖了真实的效应(Jia et al., 2018)。错误解读不显著结果也可能是出版偏倚的原因(Franco et al., 2014; Kühberger et al., 2014), 由此可能诱发研究者的 p 值操纵(p-hacking)行为(Head et al., 2015), 从而导致研究难以重复或者效应量严重减小(Baker, 2016; Klein et al., 2014; Open Science Collaboration, 2015; 胡传鹏 等, 2016)。因此, 研究者在科学研究过程中应加强对不显著结果解读的严谨性, 避免带来消极后果。

5 总结

通过分析 5 本中文心理学期刊上的 500 篇实证研究, 本研究发现中文文献中阴性陈述较为普遍, 且比例高于国际期刊, 表明不显著结果在心理学实证研究中有重要的地位。而在不显著结果的解读方面, 中文期刊中的错误解读比例小于国际期刊中的比例。另外, 贝叶斯因子的分析表明文献中不显著结果的数据并不能提供较强的支持零假设的证据。总的来说, 国内研究者需要进一步加强对不显著结果的认识, 并使用恰当的统计方法来评估数据对零假设的支持程度, 以减少对不显著结果的错误解读, 提高心理学研究的质量。

致谢: 感谢阿姆斯特丹大学心理学系的孙睿博士和斯坦福大学心理学系的赵轩博士对本文英文摘要的校读与反馈。

参考文献

程开明, 李泗娥. (2019). 科学研究中的 P 值: 误解、操纵

及改进. *数量经济技术经济研究*, (7), 117–136. doi: 10.13653/j.cnki.jqte.2019.07.007
崔诣晨, 王沛, 崔亚娟. (2019). 知觉冲突印象形成的认知控制策略: 以刻板化信息与反刻板化信息为例. *心理学报, 51*(10), 1157–1170. doi: 10.3724/SP.J.1041.2019.01157
郝丽, 刘乐平, 申亚飞. (2016). 统计显著性: 一个被误读的 P 值. *统计与信息论坛, 31*(12), 3–10.
胡传鹏, 孔祥祯, Eric-Jan Wagenmakers, Alexander Ly, 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展, 26*(6), 951–965. doi: 10.3724/SP.J.1042.2018.00951
胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题: 从危机到契机. *心理科学进展, 24*(9), 1504–1518. doi: 10.3724/SP.J.1042.2016.01504
陆春雷, 王珺, 宋琼雅, 贾彬彬, 许岳培, 胡传鹏. (2020). *从不显著结果中提取信息的方法: 原理及其实现*. 2020-10-21 取自 www.chinaxiv.org/abs/202001.00113
骆大森. (2017). 心理学可重复性危机两种根源的评估. *心理与行为研究, 15*(5), 577–586.
吕小康. (2012). Fisher 与 Neyman-Pearson 的分歧与心理统计中的假设检验争议. *心理科学, 35*(6), 1502–1506. doi: 10.16719/j.cnki.1671-6981.2012.06.042
吕小康. (2014). 从工具到范式: 假设检验争议的知识社会学反思. *社会, 34*(6), 216–236. doi: 10.15992/j.cnki.31-1123/c.2014.06.011
卢淑华. (2009). *社会统计学*(第四版). 北京: 北京大学出版社.
谢书书, 张积家, 朱君. (2019). 颜色范畴知觉效应发生在大脑两半球: 来自纳西族和汉族的证据. *心理学报, 51*(11), 1229–1243. doi: 10.3724/SP.J.1041.2019.01229
张厚粲, 徐建平. (2015). *现代心理与教育统计学*(第四版). 北京: 北京师范大学出版社.
仲晓波. (2016). 关于假设检验的争议: 问题的澄清与解决. *心理科学进展, 24*(10), 1670–1676. doi: 10.3724/SP.J.1042.2016.01670
Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PloS One, 12*(8), e0182651. doi: 10.1371/journal.pone.0182651
Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., … Wagenmakers, E. -J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 357–366. doi: 10.1177/251524591877374
Algermissen, J., & Mehler, D. M. (2018). May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology, 119*(6), 2114–2117. doi: 10.1152/jn.00765. 2017
American Psychological Association. (2010). *Publication Manual of the American Psychological Association*. Washington DC: American Psychological Association.
Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305–307. doi: 10.1038/d41586-019-00857-9
Baker, M. (2016). 1, 500 scientists lift the lid on reproducibility. *Nature, 553*, 452–454. doi: 10.1038/533452a
Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. doi: 10.1038/nrn3475
Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science, 2*(3), 233–239. doi: 10.1177/2515245919858072
Chen, X., Lu, B., & Yan, C. -G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping, 39*(1), 300–318. doi: 10.1101/128645
Chuard, P. J., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLoS Biology, 17*(1), e3000127. doi:10.1371/journal.pbio.3000127
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. doi: 10.3389/fpsyg.2014.00781
Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89. doi: 10.1016/j.jmp.2015.10.003
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. doi: 10.1037/h0044139
Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891–904. doi: 10.1007/s11192-011-0494-7
Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper. *Perspectives on Psychological Science, 7*(6), 661–669. doi: 10.1177/1745691612462587
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. doi: 10.1037/h0031619
Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. doi: 10.1126/science.1255484

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (R package version 0.84.1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=irr.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. SAGE Publications, Inc. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 392–410). Thousand Oaks, CA: SAGE Publications, Inc.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. doi: 10.1007/s10654-016-0149-3

Gronau, Q. F., Ly, A., & Wagenmakers, E. -J. (2019). Informed bayesian t-tests. *The American Statistician*, 1–14. doi: 10.1080/00031305.2018.1562983

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3). doi: 10.1371/journal.pbio.1002106

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E. -J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PloS One, 13*(4), e0195474. doi: 10.1371/journal.pone.0195474

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Jia, X. -Z., Zhao, N., Barton, B., Burciu, R., Carriere, N., Cerasa, A., … Zang, Y. -F. (2018). *Small effect size leads to reproducibility failure in resting-state fMRI studies*. Retrieved October 21, 2020, from https://www.biorxiv.org/content/10.1101/285171v1

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods (5th ed.)*. London, England: Edward Arnold.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. doi: 10.1027/1864-9335/a000178

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. doi: 10.1177/1745691611406925

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review, 25*(1), 178–206. doi: 10.3758/s13423-016-1221-4

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS One, 9*(9), e105825. doi: 10.1371/journal.pone.0105825

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*. Advance online publication. http://doi/10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. doi: 10.1177/2515245918770963

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. doi: 10.2307/2529310

Lee, M. D., & Wagenmakers, E. -J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q. F., & Wagenmakers, E. -J. (2018). Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science, 1*(3), 367–374.doi: 10.1177/2515245918779348

Ly, A., Verhagen, J., & Wagenmakers, E. -J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology, 72*, 43–55. doi: 10.1016/j.jmp.2016.01.003

Ly, A., Verhagen, J., & Wagenmakers, E. -J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32. doi: 10.1016/j.jmp.2015.06.004

Lyu, X. -K., Xu, Y. P., Zhao, X. -F., Zuo, X. -N., & Hu, C. -P. (2020). Beyond psychology: The prevalence of misinterpretation of p-value and confidence intervals across different fields. *Journal of Pacific Rim Psychology, 14*, e6. doi: 10.1017/prp.2019.28

Lyu, Z. Y., Peng, K. P., & Hu, C. -P. (2018). P-value, confidence intervals and statistical inference: A new dataset of misinterpretation. *Frontiers in Psychology, 9*, 868. doi: 10.3389/fpsyg.2018.00868

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Virginia Beach, VA: Chapman and Hall/CRC Press.

Meehl, P. E. (1967). Theory-testing in psychology and

physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. doi: 10.1086/288135

Miller, G. (2011). ESP paper rekindles discussion about statistics. *Science, 331*(6015), 272–273. doi: 10.1126/science.331.6015.272

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). *BayesFactor: Computation of Bayes factors for common designs* (Version 0.9.12-2) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/BayesFactor/BayesFactor. pdf.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. doi: 10.1037/1082-989X.5.2.241

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943–950. doi: 10.1126/science.aac4716

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553. doi: 10.1037/0033-2909.113.3.553

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Schäfer, T., & Schwarz, M. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813. doi: 10.3389/fpsyg.2019.00813

Schönbrodt, F. (2015). *Grades of evidence – A cheat sheet* [Web log post]. Retrieved from http://www.nicebread.de/grades-of-evidence-a-cheat-sheet/.

Signorell, A. (2017). *DescTools: Tools for descriptive statistics* (Version 0.99.22) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/DescTools/index.html.

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147*(6), 905–923. doi: 10.1037/xge0000424

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. -J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician, 72*(4), 303–308. doi: 10.1080/00031305.2016.1264998

Wagenmakers, E. -J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example Applications with JASP. *Psychonomic Bulletin and Review, 25*(1), 58–76. doi: 10.3758/s13423-017-1323-7

Wagenmakers, E. -J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. doi: 10.1037/a0022790

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. doi: 10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. -J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. doi: 10.1177/1745691611406923

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance*. Ann Arbor: University of Michigan Press.

Interpreting nonsignificant results: A quantitative investigation based on 500 Chinese psychological research

WANG Jun[1], SONG Qiongya[1], XU Yuepei[2,3], JIA Binbin[4], LU Chunlei[5], CHEN Xi[6], DAI Zixu[7], HUANG Zhiyue[8], LI Zhenjiang[9], LIN Jingxi[10], LUO Wanying[11], SHI Sainan[12], ZHANG Yingying[13], ZANG Yufeng[14], ZUO Xi-Nian[15], HU Chuanpeng[16]

([1] *Department of Psychology, Sun Yat-Sen University, Guangzhou* 510006, *China*)

([2] *Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

([3] *Department of Psychology, Chinese Academy of Sciences, Beijing* 100049, *China*)

([4] *School of Psychology, Shanghai University of Sport, Shanghai* 200438, *China*)

([5] *College of Teacher Education, Zhejiang Normal University, Jinhua* 321000, *China*)

([6] *Person, Shanghai* 200122, *China*)

([7] *School of Psychology, South China Normal University, Guangzhou* 510631, *China*)

([8] *Tisch School of the Arts, New York University, New York* 11201, *the United States*)

([9] *School of Education, Soochow University, Suzhou* 215123, *China*)

([10] *Institute of Education Science, Heilongjiang University, Harbin* 150080, *China*)

([11] *School of Psychology and Cognitive Sciences, Peking University, Beijing* 100871, *China*)

([12] *School of Psychology and Cognitive Sciences, East China Normal University, Shanghai* 200063, *China*)

([13] *Faculty of Psychology, Southwest University, Chongqing* 400715, *China*)

([14] *Center for Cognition and Brain Disorders, Hangzhou Normal University, Hangzhou* 311121, *China*)

([15] *National Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing* 100875, *China*)

([16] *Leibniz Institute for Resilience Research, Mainz* 55131, *Germany*)

Abstract: Nonsignificant results are common in psychological research and can be easily misinterpreted as evidence for accepting null hypothesis. This misinterpretation may lead to false statistical inferences in empirical research. However, how prevalent this misinterpretation exists in Chinese published psychological studies is unknown. To answer this question, we randomly selected 500 empirical research papers published between 2017 and 2018 in *Acta Psychological Sinica*, *Journal of Psychological Science*, *Chinese Journal of Clinical Psychology*, *Psychological Development and Education*, *Psychological and Behavioral Studies*, screened articles in which the abstracts contained any sentences that indicated nonsignificant results (we call these sentences "negative statements" hereafter). We then read those articles and extracted negative-statements-related statistics and their interpretations, and evaluated the correctness of each interpretation. Finally, we calculated Bayes factors based on the available *t* values in these nonsignificant results. The protocol was pre-registered at OSF (https://osf.io/czx6f). We found that (1) out of 500 empirical research, 36% of their abstracts (n = 180) contained negative statements; (2) in those 180 articles, we extracted 236 nonsignificant results and corresponding interpretations, and found that 41% of these interpretations was incorrect, (3) Bayes factor analysis revealed that only 5.1% (n = 2) of available nonsignificant *t*-values (n = 39) can provide strong evidence in favor of null hypothesis ($BF_{01} > 10$). We compared the results with Aczel et al. (2018) and discussed the potential reasons that caused the misinterpretation. These data suggest that Chinese psychology researchers need to improve their understanding of nonsignificant results and statistical inference.

Key words: nonsignificant results, null-hypothesis significance testing, Bayes factors, meta-research